CO-DESIGNING, CO-DEVELOPING, AND CO-IMPLEMENTING AN INSTITUTIONAL DATA REPOSITORY SERVICE

Michael Witt (mwitt@purdue.edu), Purdue University

ABSTRACT

In January of 2011, the National Science Foundation began requiring that all proposals for research funding include data management plans. At the time of the mandate, Purdue University's libraries and campus information technology units had been collaborating on enhancements to the HUBzero virtual research environment. These efforts were parlayed into the development of an institutional, digital data repository and service with the support of the campus research office. In the process, local library science practices have been extended to facilitate research data curation and cyberinfrastructure on campus. Librarians are consulting on data management plans, conducting data reference and instruction, advising on data organization and description, and stewarding collections of data within an evolving library service framework.

CONTEXT: THE DATA DELUGE

The Fourth Paradigm: Data-Intensive Scientific Discovery expounds upon the paradigm shift in science presented by Jim Grey from empirical to theoretical to computational to data-driven science, which is also known as e-Science (A. J. G. Hey, Tansley, & Tolle, 2009). The subsequent adoption of cyberinfrastructure has resulted in

a “data deluge” widely reported in both scholarly literature (T. Hey & Trefethen, 2003) (Lord, Macdonald, Lyon, & Giaretta, 2004) (Gershon, 2002) and the popular press (Anderson, 2008) (Cukier, 2010). A workshop convened by the National Science Board in 2005 produced a report, “Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century,” that recognized the value of these data and began to characterize collections of data; it built a foundation for the future development of government policies to ensure the stewardship of datasets and preserve their value for improving and advancing science (NSF, 2005). Data are valuable for validating reported research findings and can be reused to advance the original research or new lines of inquiry. By preserving and sharing existing datasets, the cost of generating new data from scratch may be avoided in some cases (Witt, 2008).

An expanded role for research libraries in digital data stewardship was forecasted by an Association of Research Libraries (ARL) workshop report to the NSF in 2006 (ARL, 2006). This forecast was substantiated in August 2010 by a survey of 57 ARL libraries, of which 21 libraries reported that they currently provide infrastructure or support services for e-Science, and an additional 23 libraries are in planning stages (Soehner, Steeves, & Ward, 2010).

Among the drivers for data sharing are mandates from federal funding agencies requiring researchers to submit data management plans with their proposals for grants. Since 2003, the National Institutes of Health has required data sharing for grants over $500,000 (NIH, 2003). The National Institute of Justice requires data-archiving strategy to be submitted 90 days before the end of a funded project (NIJ, 2010). Lastly, and perhaps most significantly, the NSF included an explicit requirement for data

management plans in grant proposals effective on January 18, 2011 (NSF, 2011). Across the Atlantic, the situation in the United Kingdom is similar with the Digital Curation Centre describing data management plan requirements coming from the Research Councils (e.g., the AHRC, BBSRC, EPSRC, ESRC, MRC, NERC, and STFC), Cancer Research UK, and the Wellcome Trust (Digital Curation Centre, 2011).

A number of academic and research libraries are beginning to take a more active role in data management on their campuses, applying library science principles to help address the data deluge. This includes a wide range of activities such as assisting researchers formulate funder-required data plans, adapting library practice to help organize and describe research datasets, developing data collections and data repositories, digital preservation, and data literacy. In some cases, librarians are extending their wealth of knowledge and experience gained from three decades of social science data librarianship to other disciplines. Some are adapting instruction and reference approaches to directly address data needs, for example, by offering data literacy and data reference—helping patrons find data and integrate it into their learning, teaching, and research. This paper describes the process of collaboration that helped produce new data services at Purdue University that were shaped by the university's library and information technology units collaborating on a series of developments to the HUBzero Platform for Scientific Collaboration[1].

AN OVERVIEW OF HUBZERO

[1] http://hubzero.org

In 2002, the NSF-sponsored Network for Computational Nanotechnology (NCN) began development of nanoHUB.org as a web platform to foster a virtual community of nanotechnologists by enabling them to develop, execute, and share simulation tools online. nanoHUB.org was developed on an open-source LAMP stack (Linux, Apache, MySQL, PHP) and utilizes the Joomla content management system to support the submission and sharing of a wide variety of digital content such as tutorials, online presentations, animations, videos, and papers. In addition to rich content, the nanoHUB supports a suite of collaborative functionality including tagging and annotation, ranking, wikis, calendars, citation tracking, and a job board. The Rappture toolkit enables programmers to easily create or port software to run within a web browser from the HUB, which also supports shared desktops via remotely accessible virtual machines and access to backend computational and storage resources on the TeraGrid and Open Science Grid. By 2007, nanoHUB.org hosted over 1,000 resources that were accessed from 172 different countries (Klimeck, McLennan, Brophy, Adams, & Lundstrom, 2008). nanoHUB.org is on track to exceed 200,000 total users by the end of 2011 ("nanoHUB.org - Usage: Overview," 2011) and is regularly cited by the NSF and others as a cyberinfrastructure success story.

With subsequent funding from NSF, the nanoHUB was retooled into the "HUBzero Platform for Scientific Collaboration" and made available for implementations for other scientific communities. The non-profit HUBzero Consortium was established to guide and sustain the development and support of HUBzero, which was released as open source software in April 2010. Over 25 "hubs" have been launched and provide virtual research environments to a wide diversity of communities such as

earthquake engineering, clinical and translational research in healthcare, manufacturing techniques, STEM education, assistive technology, National Parks rangers, and research ethics.

COLLABORATING ON CYBERINFRASTRUCTURE

A series of collaborations between campus IT, Information Technology at Purdue (ITaP), and the Libraries began on various HUB-related projects as early as 2006. Motivated by a desire to present HUB content in a more scholarly context, NCN consulted with the Libraries on standards for metadata and information architecture. These discussions led to NCN supporting a graduate research assistant programmer in the Libraries to develop an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[2] data provider for HUBzero to enable scholarly search engines and other service providers to harvest metadata records that describe content in the HUB. The data provider was written in PHP and included a mapping of data from the backend MySQL database to Dublin Core, serialized using eXtensive Markup Language (XML). The next year, another graduate student assistant was funded to implement OpenURL Content Objects in SPAN (COinS)[3] to expose structured citation metadata to scholarly search engines such as Google Scholar and to enable users of citation management software like Zotero to easily download formatted citations. HUB objects are also presented with example citations (e.g., "Cite this work as follows") to encourage users to cite their use of them. Investigating solutions for exposing citations raised the issue of unique identifiers and persistent links to HUB objects. Again, a graduate student programmer was funded

[2] http://www.openarchives.org/pmh
[3] http://ocoins.info

by ITaP and NCN to work in the Libraries to survey options for persistence that existed at that time such as ARKs, URNs, and PURLs. A recommendation was made to implement the Handle system[4], and software was developed to integrate nanoHUB.org with Handle to generate persistent identifiers for simulation tools. Other collaborations involved librarians interacting with particular HUB communities and consulting on their information organization and description. In one case, a librarian and data research scientist from the Libraries analyzed the content of the Center for Assistive Technology's HUB and created a controlled vocabulary and classification scheme for it (Carlson & Yatcilla, 2010).

Opportunities to pursue external grants motivated further collaboration. With support from a National Leadership Grant from the Institute of Museum and Libraries Services, Clemson University, Purdue, and the National Parks Service began development of the Open Parks Grid. A librarian from Purdue is a co-principal investigator on the grant with the HUBzero Project director from ITaP serving as an advisor. This work has resulted in the integration of HUBzero with the Semantic Web using Linked Data[5] and vocabularies such as the Open Archives Initiative Object Reuse and Exchange (OAI-ORE)[6].

In addition, organizing responses from the campus to the NSF's two DataNet calls in 2008 and 2009 for grant proposals from the Office of Cyberinfrastructure brought together a broad group of librarians, computer engineers and scientists, information technologists, and domain scientists. These meetings stimulated thinking about research

[4] http://handle.net
[5] http://linkeddata.org
[6] http://www.openarchives.org/ore

data management at the institutional level, and a campus-wide group was subsequently convened by the Vice President for Research that was chaired by the Dean of Libraries and the Vice Presidents of Information Technology and Research. The group included broad representation of faculty from different departments who were engaged in data-intensive research. Beginning in May 2010, dialog from the monthly meetings of the faculty group highlighted the need of our researchers for help with research data management and a plethora of differences in research practices, norms, and expectations related to data from department to department and researcher to researcher. These meetings concluded in August 2010 with a report submitted to the OVPR.

THE PURDUE UNIVERSITY RESEARCH REPOSITORY

Through the fall and winter of 2010, a steering committee made up of the Dean of Libraries, the Vice President of Information Technology (CIO), and Vice President for Research used the report and experience of the faculty meetings to pursue the creation of a campus resource for research data management. In March 2011, the steering committee created the Purdue University Research Repository (PURR) Working Group to bring faculty and staff from the units who had been doing independent work (Libraries, ITaP, Sponsored Program Services, and the OVPR) in this area together with a charge to define and deploy an institutional data repository service using the HUBzero software. Three years of HUBzero hosting and support were purchased with the cost divided evenly among the three partners. The Working Group was chaired by the Libraries' Interdisciplinary Research Librarian and included representation from the three units plus the Sponsored Program Services:

- Associate Vice President for Research, OVPR
- Associate Dean for Digital Programs and Information Access, Libraries
- Associate Dean for Research, Libraries
- Data Services Specialist, Libraries
- Chemical Information Specialist, Libraries
- Assistant Director of Pre-Award Services, Sponsored Programs Services
- Managing Director, Launching Centers and Institutes, OVPR
- Visiting Assistant Professor, Libraries
- HUBzero Project Director, ITaP
- University Archivist, Libraries
- HUB Community Liaison, ITaP

The composition of the group represented a collaboration of stakeholders among the service providers in the university. The research office (OVPR) is invested in fostering an environment of compliance with funder requirements such as the NSF mandate as well as helping investigators submit more competitive proposals. Because the data management plan may be reviewed as a part of the proposal, a good plan may improve reviewer scores. The OVPR drew upon their wealth of experience with researchers, funders, and policy, translating and incorporating their needs into the design of PURR. Sponsored Program Services helps investigators prepare proposals and performs grant administration. They closely monitored proposal submissions and awards and gave valuable, real-time information and feedback on the constitution of data management plans and the rates of adoption and success. Information technology professionals and research computing specialists at ITaP had expertise and capacity to tackle challenges

related to technology such as server and storage infrastructure. The HUBzero platform was selected for PURR mainly because it was developed at Purdue and offered much of the desired functionality with a large, local base of support staff and software developers.

ITaP set up an instance of HUBzero as a prototype of PURR and demonstrated its functionality to the Working Group, who began meeting for an hour, every other week. The group used the prototype as its primary means of collaborating online in between meetings—essentially using PURR to develop PURR and giving themselves the experience of being a user of the system. This led to everyone on the group gaining a familiarity with the platform and offering immediate feedback for enhancing the system. The Working Group created a private project space with a wiki for collaborating on agendas, recording minutes, co-creation and editing of documents, and publishing resources.

Full participation was encouraged in meetings, with all members able to put items on the agenda for discussion or use the whiteboard to brainstorm and diagram ideas. Early meetings focused on creating a high-level definition of the PURR service:

> "PURR provides an online, collaborative working space and data-sharing platform to support the data management needs of Purdue researchers and their collaborators. It is an initiative of the Purdue University Libraries, Information Technology at Purdue, and the Office of the Vice President for Research. PURR is being developed into a Trustworthy Digital Repository that uses DataCite Digital Object Identifiers (DOIs) and other standards to support the discovery,

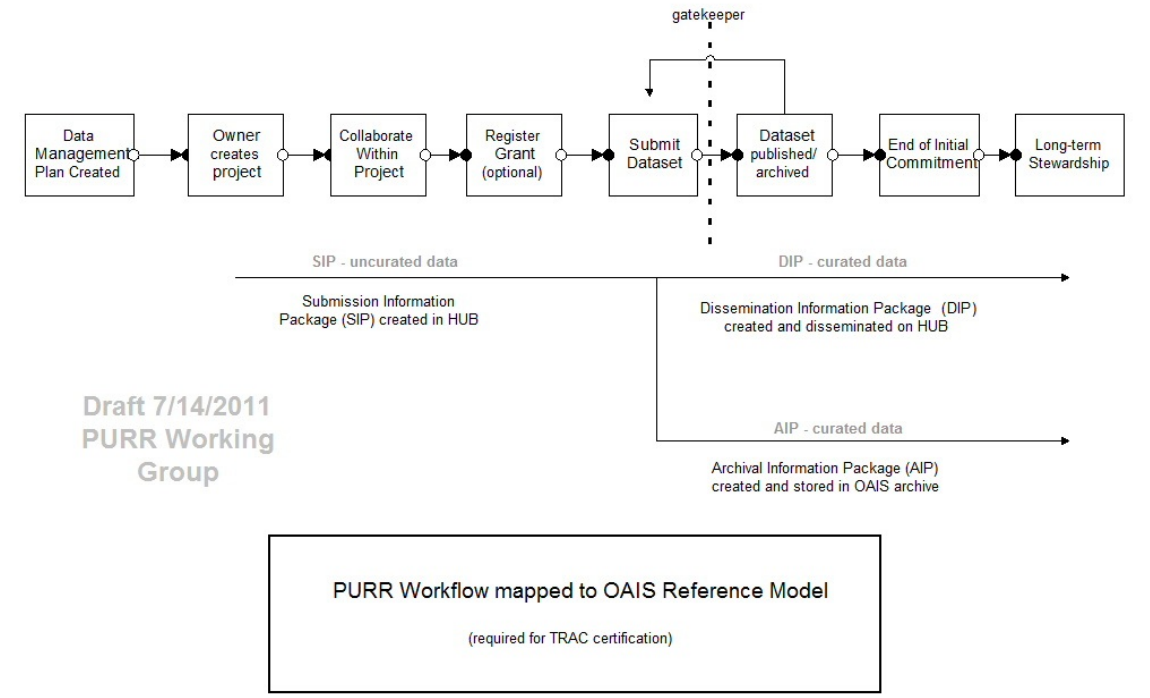
> use, and long-term preservation of data. PURR is based on the HUBzero platform." (Purdue University, 2011a)

The Working Group brainstormed and formulated a high-level workflow for PURR that revolves around projects. A project is a private, dedicated working space on PURR for users to collaborate and prepare data for publication and curation for a research project or study. It includes a small, default storage allocation for uploading and sharing files, a wiki, to-do list, and other collaborative tools. After creating a project, the project owner can invite collaborators from Purdue or other institutions to join the project via a link sent by email. Any Purdue employee can create a project. Ownership of projects can be extended to other users; however, all projects must maintain at least one owner who is a Purdue employee. Projects and their associated working spaces are transient and expire after a defined period of time; however, datasets can be published from within the projects to be made publicly available or preserved in a dark archive for longer periods of time. These datasets are referred to and maintained as "curated data." Project owners can register a grant award with their project and receive a larger storage allocation and longer project duration.

This service definition and high-level workflow were presented for approval to the PURR Steering Committee along with cost information for three different options for supporting storage. The committee decided to provision 500 megabytes of project storage (i.e., "working space") for three years by default for all projects with up to 50 megabytes of curated data. If a grant is registered and associated with a project, this allotment increases to 100 gigabytes of project space for 10 years with up to 10 gigabytes of curated data. All published and archived (i.e., "curated") datasets are maintained for at

least 10 years. Lastly, a project owner can purchase additional storage using departmental or other funds.

With a platform and high-level service definition and workflow in place, the Working Group began a series of extensive discussions to identify the steps in the workflow at a lower level, in terms of what functionality would need to be provided by the repository system, what tasks would need to be accomplished by staff, and what policies may be needed for each step. Each step was discussed and sketched on the whiteboard multiple times as well as recorded and revised as a page in the PURR project wiki. Tasks were identified and researched by voluntary sub-groups within the Working Group, who worked independently and then reported back at subsequent meetings with proposals for the group to consider and implement. This was an iterative process, adding, building, and redefining steps as progress was made. The default HUBzero software provided a baseline of functionality that enabled a quicker and easier design process, because the members of the group could articulate functional requirements for PURR, in terms of extensions to HUBzero, by modifying user interfaces and workflows that were already familiar to the group. By June 2011, the group had drafted a final diagram of the PURR workflow and loosely mapped it to the Open Archival Information System reference model.



Around this same time, the Libraries' members of the Working Group began meeting on opposite weeks of the full group's meetings in order to discuss issues that were primarily situated in the Libraries such as digital preservation, persistent identifiers, metadata, data discovery, and librarian integration. The Libraries' group worked in a similar fashion with participants identifying tasks, spinning off sub-groups to work on them, and reporting accomplishments back to the full PURR Working Group.

One of the main objectives of this group was to build opportunities for librarians to engage researchers and participate actively in data curation into the design of PURR. When a project is created, a subject specialist librarian can be assigned to it based on the department affiliation of the project owner. Department codes can be retrieved from the

online campus directory system and mapped to subjects covered by individual librarians. The appropriate subject specialist librarian is then notified by email that the project has been initiated and given the contact information of the project owner. The librarian then has the opportunity to contact and engage the group, to learn more about their research and consult or collaborate on the project. Later in the research cycle, when a project member submits a dataset for publication or archive, the librarian will be notified again and is required to approve the dataset before it is published or archived. The librarian does not evaluate the quality or veracity of the data but instead performs a series of checks to make sure the data is an appropriate submission for PURR, in an acceptable format, and includes sufficient metadata. A repository coordinator and data service specialists are available to support librarians and provide redundancy in the event one is not able to act quickly on a submission. This workflow is similar to the workflow for Purdue's institutional e-print repository, so it is familiar to faculty and staff. At some point on the horizon, the initial 10-year commitment to maintain the published or archived dataset expires. The project owners will receive an email 6-12 months before this occurs, and if they do not purchase additional storage the dataset is remanded to the Libraries. The librarian who is associated with the project is notified and can evaluate the dataset for inclusion in the regular library collection. If the dataset is selected for the collection, the Libraries maintains it as a function of its collection management. If it is not selected for the collection, the dataset is removed and its identifier is updated to resolve to a de-accession notice.

Software development was led by ITaP with iterative updates and functional requirements communicated between the software developers and the Working Group by

the HUB Community Liaison. Many of the features developed for PURR will also be contributed to the HUBzero open source project, and PURR benefitted from the newest features being developed by others thanks to the involvement of the HUBzero Project Director in the group. Development of the basic repository functionality and workflow for PURR was completed in December 2011.

DATA MANAGEMENT PLANNING

In January of 2011, the NSF data management plan mandate went into effect, requiring that all grant proposals submitted to them be accompanied by a two-page supplement that describes how the investigator will disseminate and share the results of their research. The NSF Grant Proposal Guide suggests that such plans may include the type of data that will be produced in the research; what standards for description and format will be used; policies for sharing data that address intellectual property, privacy, and rights, provisions for reuse of data; and plans for archiving and preservation of data (NSF, 2011). Purdue tracked the development of the mandate closely as the NSF is the largest federal sponsor of research on its main campus with over $100 million in grant awarded annually (Purdue University, 2011b). By the time the mandate arrived, the Libraries had established a reputation on campus for expertise in data curation through the advocacy of its Dean of Libraries and the research and work of its Distributed Data Curation Center[7] and affiliated librarians. The Libraries had included institutional data curation in its strategic plan as early as 2006. When the plan was updated in 2011, the Libraries further challenged itself to "lead in data-related scholarship and initiatives" as a function of facilitating scholarly communication as well as to "lead in international

[7] http:/d2c2.lib.purdue.edu

initiatives in information literacy and e-science and utilize [its] expertise in the provision of information access, management, and dissemination to collaborate on campus-wide goals" (Purdue University Libraries, 2011). Consideration of data and e-science is woven into the current plan's goals and objectives, such as the inclusion of data in information literacy, the identification and building of collections that are unique to Purdue, and the development and promotion of new publishing models (Purdue University Libraries, 2011).

Thus, the OVPR turned to the Libraries to collaborate in raising awareness of the new mandate and educating researchers about it. The Libraries' Data Services Specialist led an ad-hoc group of librarians and research office staff to organize a series of data management plan workshops that were hosted by the OVPR and promoted by both the libraries and research offices: two in the spring and two in the fall. The workshops presented an overview of the mandate and what tools and services are available to help researchers meet it, including PURR. There were speakers from the Libraries, ITaP, and the OVPR, and the most recent workshop was video-recorded and uploaded to PURR to be archived and viewed on-demand[8].

In conjunction with the PURR Working Group, the Libraries developed a series of supporting materials for the workshop and for general use that were made available on PURR. A "Data Management Plan Overview" provides investigators with a concise list of questions that begin to address issues that were derived from research performed as a part of the Data Curation Profiles that was supported by the Institute of Library and

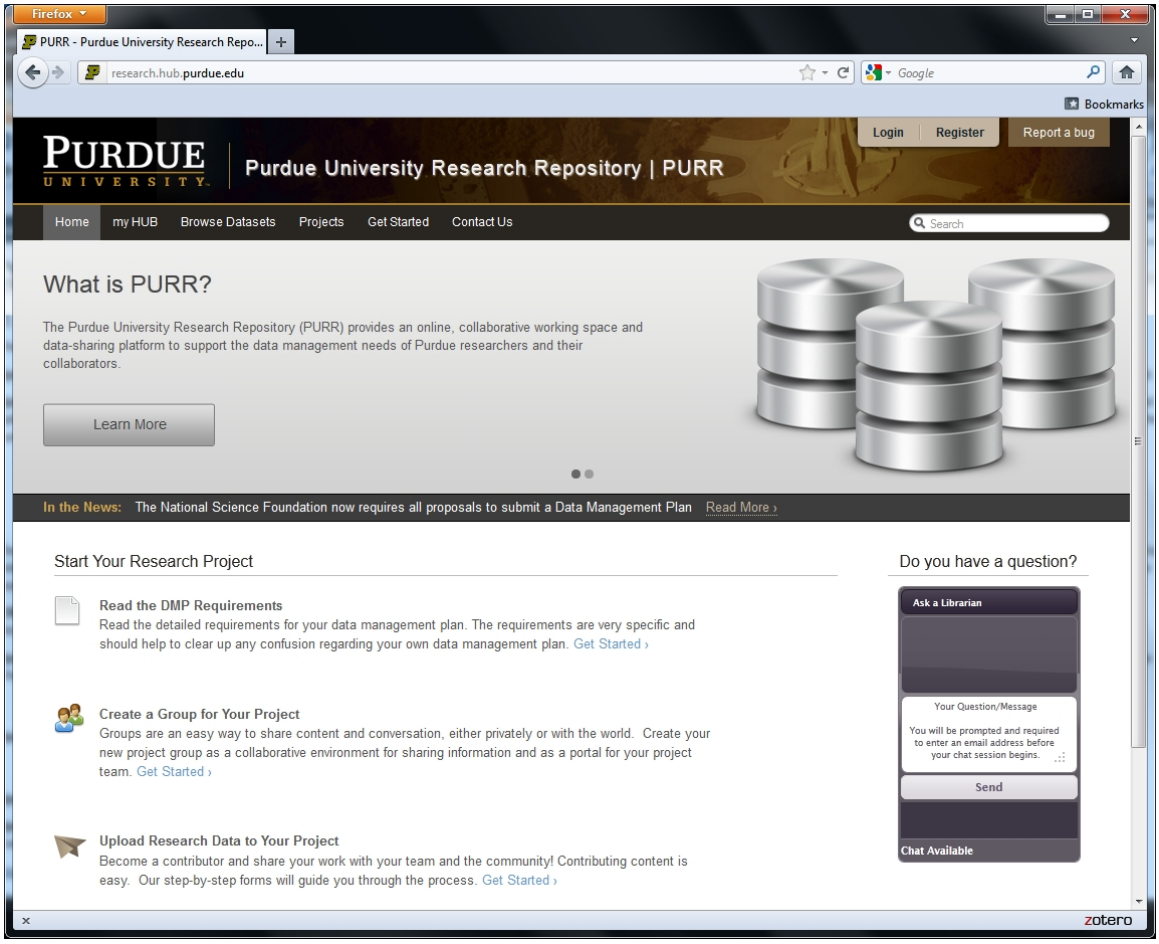
[8] https://research.hub.purdue.edu/resources/16/download/2011.09.23-McLennan-DMPworkshop-640x360.mp4

Museum Services (Witt, Carlson, Brandt, & Cragin, 2009). A longer form of the overview, the "Data Management Plan Self-Assessment" was written by Jacob Carlson and produced by the Libraries and OVPR to lead investigators through a set of questions that, in answering, will provide them with the basic building blocks of a data management plan (Purdue University, 2011c). A set of example data management plans are provided from past projects from a variety of funding agencies as well as links to data management planning tools. An extensive tutorial was created for designing data management plans for the NSF specifically, and more funder-specific tutorials are planned. Other helpful information is available in a Frequently Asked Questions (FAQ) format in the PURR Knowledge Base[9], for which the Libraries, OVPR, and ITaP collaborate on answering.

A general boilerplate text[10] is provided to investigators who intend to use PURR as a part of the data management plan. In one long paragraph, the boilerplate describes the PURR service at a high level including its generic functionality, policies, metadata and preservation support, use of standards such as DOIs, and its progress towards establishing itself as a trustworthy digital repository. Because all Purdue investigators can receive a default allocation of resources on PURR, the boilerplate can include an institutional commitment of data management resources to the proposed project by default. Lastly, the boilerplate links to the PURR website for more information. The boilerplate text has proven to be a very popular option; Sponsored Programs Services scanned and identified PURR as a component of the data management plans of 34% of

[9] http://research.hub.purdue.edu/kb/AboutPURR
[10] http://research.hub.purdue.edu/dmp/usehub

proposals submitted to the NSF from Purdue for the first ten months since the mandate, the majority of which utilized the boilerplate text.



The PURR website, workshops, and other materials have resulted in many researchers contacting librarians for assistance with data management and collaboration on data-intensive research. Consultations have taken place in-person, over email, and on the telephone, and the involvement of librarians has ranged from helping to create a plan from scratch to reviewing plans, writing letters of support for grant proposals, and being named on grants as co-principal investigators and senior personnel.

DATA REFERENCE

Data reference is not a new concept; in fact, it has been a part of the regular practice of librarianship, especially in the social sciences, for decades. The community of librarians and information professionals that have evolved around the International Association for Social Science Information Services and Technology (IASSIST) and the Interuniversity Consortium for Political and Social Research (ICPSR) repository service are exemplars. The underlying principals of library science that have been traditionally applied to print literature can also be applied to scientific data. Much like a librarian who is new to a particular subject, he or she has the training to organize, develop, and enhance the use of a previously unfamiliar collection (Mullins, 2011). Librarians can approach the landscape of data using the same tools: learning what data are important for scholarship in their areas, how and where these data are described and stored, and interpreting relevant issues in the context of the data (e.g., intellectual property, preservation, metadata, authenticity, etc.) Librarians who are subject specialists can specialize and incorporate data into their collection, instruction, and reference activities for patrons in their subject areas.

Like most academic and research libraries, Purdue offers digital reference services to patrons using an instant messenger chat and email widget that can be embedded into the library's web pages. The QuestionPoint software[11] that supports this service also provides an online system for routing and reporting reference transactions. The service is staffed on an hourly basis during business hours for chat reference and rolls over to email when staff are not available to chat. The digital reference service has become very popular with over 5,500 transactions logged in 2010. In the same period of

[11] http://www.questionpoint.org

time, 52 faculty and staff answered digital reference questions by email, and 31 worked one-hour shifts to answer questions via chat. New reference workers go through a formal orientation process when they begin and subsequently refresh their skills every year with mandatory, annual training sessions.

It seemed logical to extend the Libraries' existing service framework for digital reference to PURR, although the process was not as simple as copying-and-pasting the QuestionPoint widget into the PURR web pages. It was important to visualize how the service would be offered and map to the existing service as well as to anticipate what kinds of questions might be asked. The Digital Curation Center's Data Curation Lifecycle[12] was used as a basis for brainstorming potential reference questions. Even if the library or university is not equipped to address the needs represented by the questions with institutional solutions, it is appropriate that librarians apply their skills to help patrons effectively find, evaluate, and use data sources and services—even if these issues are new to them.

[12] http://www.dcc.ac.uk/resources/curation-lifecycle-model, image used with permission.

After consulting with the Digital Reference Coordinator and her team, a proposal to extend digital reference service to PURR was submitted for approval to the Libraries' Planning and Operations Council, which includes library administrators and a broad representation of the units and divisions of the Libraries. The proposal was approved with a recommendation that a small group of designated "data librarians" triage questions with reference workers and subject librarians who may not yet be prepared to address questions about data on their own. The Libraries' Associate Dean for Research and the Data Education Working Group identified seven librarians to fill this role, and procedures were drafted to ensure the proper identification and handling of reference questions related to PURR or data management for both reference workers and the data librarians.

When a new question is entered into the system by chat or email, the reference worker tries to identify if it is a question about PURR or data management by reading the text of the question and checking the referrer URL to see if the question is coming from

the PURR website. If it is a question about PURR or data management, he or she executes a script that explains to the patron that their question will be referred to a data librarian who will get back to them in 1-2 business days and asks if this is acceptable. The question is then routed to a preset group called Data Librarians that also sends an email notice to the seven designated librarians. The first data librarian to respond to the question re-assigns it to himself or herself, to avoid confusion and duplication of effort. The data librarian contacts the appropriate subject librarian or librarians and works with them to contact the patron and address their question, noting the course of action and resolution in the system.

Handouts were created for our faculty and staff that were incorporated into the annual training session along with a hands-on activity to reinforce the new procedure, which went into effect in August 2011. As with any change, minor challenges were encountered. Two or three questions were improperly or incorrectly answered by reference workers instead of being routed to data librarians. Bookmarks were created by the Digital Reference Coordinator and sent to reference desks. The bookmarks serve a dual-purpose of raising patrons' awareness of the PURR service and reminding reference workers to anticipate questions about data. The Libraries worked with ITaP to help characterize questions that reported bugs in the system or require other technical support in order to refer those questions to them. Likewise, questions related to proposal development and administration that do not pertain to data management are referred to Sponsored Program Pre-Award Services.

While it is too early to measure the impact and success of extending digital reference to PURR, questions answered by data librarians are being tagged for future

reporting and analysis. Having the digital reference chat widget on all of the main PURR web pages ("Do you have a question? Ask a Librarian") connects librarians with users at their point of need—users who otherwise may not have considered or known that a librarian could help them with their data.

DATA DISCOVERY AND DIGITAL PRESERVATION

Librarians have implemented descriptive, technical, and administrative metadata for data objects that are managed by PURR to support a basic level of functionality such as searching and browsing datasets, maintaining relationships and semantics of files within datasets, and archiving them. The native metadata records that describe datasets have been mapped to Dublin Core with the intention of harvesting them using the OAI-PMH data provider that was previously developed. The harvested metadata will be indexed by the next iteration of the Libraries' online catalog, Ex Libris' Primo, so that research datasets can be searched and discovered alongside books, journals, and other library collections.

In 2010, Purdue University became a founding member of an international, non-profit organization, DataCite, that established a global Digital Object Identifier (DOI) registration agency for research datatsets. DataCite DOIs create unique and persistent identifiers that facilitate data citation and can be dereferenced to provide access to datasets, even if they are moved from one server to another (Brase, 2009). By 2011, DataCite had registered over one million datasets with DOIs (Farquhar, 2011). DOIs use the same technical architecture as the Handle system, and the prior experience gained and relationship formed in integrating Handle and HUBzero contributed to a rapid integration

of DataCite with PURR. All published and archived datasets in PURR receive DOIs, a value that is highlighted in the data management plan boilerplate text.

The current practices and expertise of the Libraries' Archives and Special Collections are being extended to research data. A minimal bit-level digital preservation strategy has been adopted pending the review and implementation of a new and comprehensive PURR Digital Preservation Policy that has been drafted by the working group and submitted to the Libraries' Planning and Operations Council for approval. The working group used the Trustworthy Repository Audit Checklist (TRAC) as a guiding document in their design of PURR, and it helped the group think holistically about the PURR service as robust preservation repository that can be trustworthy. TRAC outlines 84 criteria to be met by the principles of documentation, transparency, adequacy, and measurability in three sections: Organizational Infrastructure, Digital Object Management, and Technologies, Technical Infrastructure, and Security (Online Computer Library Center & Center for Research Libraries, 2007). TRAC recently went through the standardization process and has become ISO 16363[13]. The PURR Steering Committee committed up-front to building PURR as a trustworthy digital repository, which empowered the working group to use TRAC as an input to the design process and lend clarity to many functional requirements and much documentation produced by the group for PURR (e.g., mission statement, policies, job descriptions, business plan, etc.)

CONCLUSION

[13] http://public.ccsds.org/publications/archive/652x0m1.pdf

In November 2011, the PURR Working Group submitted a four-year budget and development plan to scale and sustain PURR that includes information about staffing needs, storage and infrastructure, ISO 16363 certification, and desired new functionality and services. An evaluation and assessment is proposed for 2013 to compare the group's estimates with actual use and adjust resourcing accordingly. A Data Education Working Group has been formed within the Libraries to identify and provide training to librarians to encourage and support their engagement and outreach related to data. This group has hosted a series of seminars and produced a LibGuide, "Supporting Information for Data Services,"[14] as a resource for librarians. Some librarians have begun to produce similar guides for users related to data issues (e.g., data citation[15]) and are incorporating data into their information literacy instruction. New services and extensions of existing services to address data curation are continuing to be developed. Working with data will become a mature component of librarianship when it is accepted into regular library practices: when terms like "data reference" become simply "reference" and datasets are not given any specific or specialized treatment than other library collections. These new services and the accomplishment of establishing PURR would not have been possible without the collaboration of the units involved; any future services and enhancements will be built upon this foundation of collaboration.

REFERENCES

[14] http://guides.lib.purdue.edu/dataservices
[15] http://guides.lib.purdue.edu/datacitation

Anderson, C. (2008). The end of theory: the data deluge makes the scientific method obsolete. Wired Magazine 16.07.

ARL. (2006). To Stand the Test of Time: Long-Term Stewardship of Digital Data Sets in Science and Engineering. Retrieved from http://www.arl.org/bm~doc/digdatarpt.pdf

Brase, J. (2009). DataCite - A Global Registration Agency for Research Data. Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09 (pp. 257-261). Presented at the Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09, IEEE. doi:10.1109/COINFO.2009.66

Carlson, J., & Yatcilla, J. K. (2010). The Intersection of Virtual Organizations and the Library: A Case Study. The Journal of Academic Librarianship, 36(3), 192-201. doi:10.1016/j.acalib.2010.03.001

Cukier, K. (2010, February 25). Technology: The data deluge. The Economist, 394(8671). Retrieved from http://www.economist.com/node/15579717

Digital Curation Centre. (2011). Funders' Data Policies. Retrieved December 14, 2011, from http://www.dcc.ac.uk/resources/policy-and-legal/funders-data-policies

Farquhar, A. (2011, August 24). Fostering Data Citation and Reuse: Shaping the DataCite Roadmap. Presented at the DataCite Summer Meeting, Berkeley, California, USA. Retrieved from http://datacite.org/node/41

Gershon, D. (2002). Dealing with the data deluge. Nature, 416(6883), 889-891. doi:10.1038/416889a

Hey, A. J. G., Tansley, S., & Tolle, K. M. (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Redmond, WA: Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Hey, T., & Trefethen, A. (2003). The Data Deluge: An e-Science Perspective, 809-824. doi:10.1002/0470867167.ch36

Klimeck, G., McLennan, M., Brophy, S. P., Adams, G. B., & Lundstrom, M. S. (2008). nanoHUB.org: Advancing Education and Research in Nanotechnology. Computing in Science & Engineering, 10(5), 17-23. doi:10.1109/MCSE.2008.120

Lord, P., Macdonald, A., Lyon, L., & Giaretta, D. (2004). From data deluge to data curation. Proc 3th UK e-Science All Hands Meeting (pp. 371–375).

Mullins, J. L. (2011, December 15). Personal Interview With James L. Mullins.

nanoHUB.org - Usage: Overview. (2011, November 30). Retrieved November 30, 2011, from http://nanohub.org/usage

NIH. (2003, March 5). NIH Data Sharing Policy and Implementation Guidance. Retrieved December 14, 2011, from http://grants.nih.gov/grants/policy/data_sharing/data_sharing_guidance.htm

NIJ. (2010, January 15). Data Archiving Strategies for NIJ Funding Applicants. Retrieved December 14, 2011, from http://www.nij.gov/funding/data-resources-program/applying/data-archiving-strategies.htm#note1

NSF. (2005). Long-Lived Digital Data Collections Enabling Research and Education in the 21st Century (No. US NSF NSB-05-40). Retrieved from http://www.nsf.gov/pubs/2005/nsb0540/

NSF. (2011, January). Grant Proposal Guide. Retrieved December 6, 2011, from http://www.nsf.gov/pubs/policydocs/pappguide/nsf11001/gpg_2.jsp#dmp

Online Computer Library Center, & Center for Research Libraries. (2007). Trustworth Repositories: Audit & Certificaiton: Criteria and Checklist. Retrieved from http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying-0/trac_0.pdf

Purdue University. (2011a). Internal Documentation.

Purdue University. (2011b). Purdue Data Digest: Extramural Awards by Sponsor. Retrieved December 6, 2011, b from http://www.purdue.edu/datadigest/research/res_spon.html

Purdue University. (2011c, February 4). Data Management Plan Self-Assessment Questionnaire. Retrieved December 7, 2011, c from http://research.hub.purdue.edu/resources/7

Purdue University Libraries. (2011, May 24). Strategic Plan 2011-2016. Retrieved December 6, 2011, from http://www.lib.purdue.edu/admin/stratplans/

Soehner, C., Steeves, C., & Ward, J. (2010). E-Science and Data Support Services: A Study of ARL Member Institutions. Association of Research Libraries. Retrieved from http://www.arl.org/bm~doc/escience_report2010.pdf

Witt, M. (2008). Institutional Repositories and Research Data Curation in a Distributed Environment. Library Trends, 57(2), 191-201. doi:10.1353/lib.0.0029

Witt, M., Carlson, J., Brandt, D. S., & Cragin, M. H. (2009). Constructing Data Curation Profiles. International Journal of Digital Curation, 4(3). Retrieved from http://www.ijdc.net/index.php/ijdc/article/view/137